基于文本聚类与兴趣衰减的微博用户兴趣挖掘方法 *

秦永彬 [a, b†]，孙玉洁 [a]，魏　笑 [a]

(贵州大学 a. 计算机科学与技术学院; b. 贵州省公共大数据重点实验室, 贵阳 550025)

摘　要：微博平台隐含潜在的用户信息，通过微博数据挖掘用户兴趣具有重要的社会意义。结合用户兴趣与微博信息的特点，提出了一种文本聚类与兴趣衰减的微博用户兴趣挖掘（TCID-MUIM）方法。首先，通过基于词林的同义词合并策略弥补建模时词频信息不足的弊端；其次，利用二次 Single-Pass 不完全聚类算法将用户微博划分为多个簇，将簇合并为同一文档以弥补微博文本短小难以挖掘主题信息的问题；最后，通过 LDA 模型建模，并考虑用户兴趣随时间变化的问题，引入时间因子，将微博—主题矩阵压缩为用户—主题矩阵，获取用户兴趣。实验表明，较之传统建模方法与合并用户历史微博为同一文档的建模方法，TCID-MUIM 方法挖掘的用户兴趣主题具有更好的主题区分度，且更贴合用户的真实兴趣偏好。

关键词：微博；Single-Pass 聚类；LDA 模型；用户兴趣挖掘；兴趣衰减

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.11.0743

Microblog user interest mining based on text clustering and interest decay

Qin Yongbin [a, b†], Sun Yujie[a], Wei Xiao[a]

(*a. College of Computer Science & Technology*, *b. Guizhou Key Laboratory of Public Big Data Guizhou University*, *Guiyang* 550025, *China*)

Abstract: Microblog platform contains potential user's information, through microblog data mining microblog user interest has important social significance. On account of the characteristics of user interest and microblog information, this paper put forward a method of microblog user interest mining based on text clustering and interest decay(TCID-MUIM) . Firstly, it used the synonyms combined strategy based on Tongyici Cilin to make up for the process of modeling the lack of word frequency information. Secondly, it used the double single-pass incomplete clustering algorithm to make up the problem that the microblog text was shorter so that difficult to dig the topic information. Finally, it used the LDA model modeling, as well as considering the user's interest changes with time, by introduction of time factor compresses the microblog-topic matrix into the user-topic matrix to gain user interest. Experimental results show that compared to traditional modeling methods and the modeling methods of merger user's all history microblog as the same document, the TCID-MUIM method presented which modeling results have a higher topic's differences and closer to the user's real interest preferences.

Key words: microblog; Single-Pass clustering; LDA model; user interest mining; interest decay

0　引言

信息技术的飞速发展与互联网的广泛应用，促使了微博、微信等具有强大交互性的网络社交平台的深入应用并使其融入了人们的社会化生活。微博，作为一个以用户为主体进行信息分享的广播式社交网络平台，由于信息发布便捷、内容形式多样、名人效应等特点聚集了 2.71 亿用户[1]，微博已成为人们获取信息、交流信息的重要工具。调查显示，微博用户关注的内容倾向于基于兴趣的垂直细分领域[1]，且微博平台上 61.9%的用户只浏览、点赞、评论或转发，基本不发原创微博[2]，这意味着用户使用微博的主要目的是在微博平台上获取自己感兴趣的内容。然而微博平台存在“信息过载”问题，用户很难在海量微博信息中获取感兴趣的信息。个性化推荐通过对用户信息的挖掘有针对性地为用户推荐有效的微博信息，是解决上述问题的有效方法.在此过程中，用户兴趣挖掘是为用户进行个性化信息推荐的前提。因此，本文以微博平台的用户兴趣挖掘为研究内容。

针对微博平台用户兴趣挖掘，许多学者展开了相关研究工作。针对建模文本的选择问题，Chen 等人[3]比较了用户历史微博文本及用户粉丝的微博文本用于用户兴趣挖掘的效果，发现

收稿日期：2017-11-17；**修回日期：**2018-01-08　　**基金项目：**国家自然科学基金（61540050）；贵州省重大应用基础研究项目（黔科合 JZ 字［2014］2001）；

作者简介：秦永彬（1980-），男，山东招远人，教授，博士，主要研究方向为智慧计算与智能计算、大数据管理与应用（ybqin@gzu.edu.cn）；孙玉洁（1993-），女，硕士，主要研究方向为文本挖掘、个性化推荐；魏笑（1994-），女，硕士研究生，主要研究方向为知识图谱、数据挖掘.

基于用户历史微博文本构建的模型更能表达用户兴趣。针对微博文本长度短影响挖掘效果的问题，Abel 等人[4]通过引入外部语料（网页链接）扩展微博信息，提取能反映用户兴趣的关键字；Hong 等人[5]提出了三种主题模型训练方法，其实验证明：合并用户所有微博为一个文档用于建模相较于一条微博作为一个文档或合并相同标签的微博为一个文档进行建模，能更有效地训练主题模型。上述方法中，通过网页链接引入外部语料的方法所获取的数据并非全部与微博文本内容高度相关，会影响建模效果；而合并用户历史微博为同一文档的建模方法强行抹去微博文本边界，会使得建模后主题区分度变低。

在微博平台的热点话题与信息检索领域，刘红兵和唐晓波等人[6,7]通过 LDA 模型与文本聚类相结合的方法，将具有相似话题的微博聚集成簇，分别实现微博平台的热点话题检测及微博信息的有效检索。该方法提供了解决微博文本长度短影响挖掘效果的新思路。

在上述研究的基础上，本文提出了一种基于文本聚类与时间衰减的微博用户兴趣挖掘方法（text clustering and interest decay for microblog user interest mining，TCID-MUIM）。该方法通过同义词合并策略，弥补建模过程中词频信息不足的弊端；通过二次 Single-Pass 不完全聚类算法，将用户微博划分为多个簇，将簇合并为同一文档，以弥补微博文本短小难以挖掘主题信息的问题。考虑到 LDA 主题模型具有优秀的降维能力，用于微博等具有稀疏性特征的文本建模具有一定优势[8]，因此选择通过 LDA 模型对多个微博文本进行建模。同时，考虑到用户兴趣会随时间变化的问题，提出基于时间因子的主题矩阵压缩方法，通过记忆值将微博—主题矩阵压缩为用户—主题矩阵，更准确地表达用户兴趣主题。

1　LDA 模型及用户兴趣表达

LDA 模型[9]是一种包含文档层、词汇层和主题层的三层贝叶斯模型，它也是一种概率生成模型。在 LDA 模型中，文档的生成前提是词袋模型(bag-of-words)，其把文档看成是一系列词汇的集合，忽略文中的语法和词汇的出现顺序，使得词与词之间独立可交换。LDA 模型的基本思想是，一篇文档由一系列描述文档的主题构成，每个主题下是一系列描述主题的词汇，其生成文本的方式可以用 LDA 模型的贝叶斯网络图表示。

图 1 描述了文本集中 M 个文档的生成过程。以第 m 个文档的生成为例，其生成步骤如下：a）从参数为 β 的 Dirichlet 先验分布中抽取 K 个主题对应的词汇分布 $\Phi=\{\varphi_k\}_{k=1}^{K}$；b）从参数为 α 的 Dirichlet 先验分布中抽取文档的主题分布 θ_m；c）根据主题分布概率 θ_m，从 K 个主题中抽取一个主题 $z_{m,n}$；d）从主题 $z_{m,n}$ 对应的词汇分布 $\varphi_{z_{m,n}}$ 中抽取一个词汇 $w_{m,n}$；e）重复步骤 c)d) n 次，直到生成文档中全部的 N_m 个词。

在 LDA 模型中，$w_{m,n}$ 是可观测的数据，α 和 β 是根据经验给定的先验参数。文档的主题分布 Θ 及主题下的词汇分布 Φ 都是需要推断的未知参数，可通过变分贝叶斯期望最大化算法[9]、吉布斯采样（Gibbs sampling）[10]等参数推导方法进行参数估计。

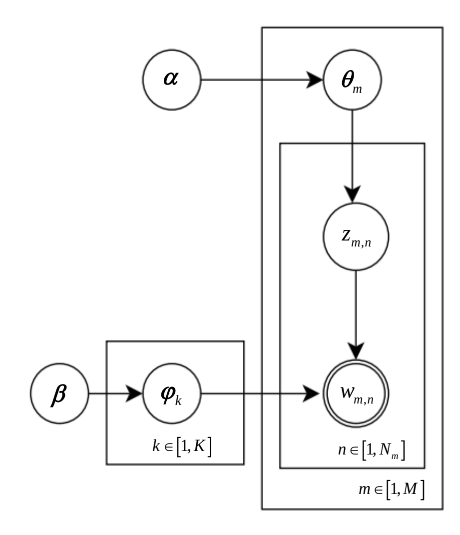


图 1　LDA 模型的贝叶斯网络图

通过 LDA 构建微博用户兴趣模型时，若输入的语料是 M 条表示为词袋模型的微博文本 W，事先给定先验参数 α、β 及需要划分的主题数 K，可训练得到用于表达用户兴趣的微博—主题矩阵 Θ、主题—词汇矩阵 Φ，如图 2 所示。

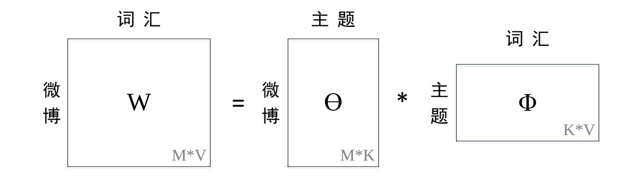


图 2　基于 LDA 模型的用户兴趣表达

2　基于文本聚类与兴趣衰减的微博用户兴趣挖掘算法

微博平台的用户兴趣挖掘可描述为：假定用户 u 的 M 条历史微博文本集 $D_u=\{d_{u1},d_{u2},\cdots,d_{uM}\}$。对用户 u 的兴趣挖掘首先是对 D_u 中的文本进行有针对性的预处理，把每条微博文本表示为特征词的集合 $W_u=\{w_{u1},w_{u2},\cdots,w_{uM}\}$；然后通过主题建模方法（如 LDA）对 W_u 进行兴趣主题建模，获取微博—主题矩阵 Θ_u 及主题—词汇矩阵 Φ_u。其中，Θ_u 描述了每条微博的主题概率分布，如图 3 所示。

主题

$$\text{微博}\quad\begin{matrix} & T_{u1} & T_{u2} & \cdots & T_{uK} \\ d_{u1} & p_{11} & p_{12} & \cdots & p_{1K} \\ d_{u2} & p_{21} & p_{22} & \cdots & p_{2K} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ d_{uM} & p_{M1} & p_{M3} & \cdots & p_{MK} \end{matrix}$$

图 3　微博-主题矩阵 Θ_u

图 3 中，Θ_u 表达了每一条用户历史微博的主题概率分布，但很难理解为用户自身的兴趣主题分布。除此之外，LDA 虽然在挖掘文本主题方面具有一定优势，但微博本身的一些固有特点还是为用户兴趣挖掘带来了以下问题：

a）针对微博文档集进行 LDA 建模时，需要预先设定主题数 K。K 的选取对于兴趣建模效果具有很大影响，K 的过小和过大都会导致主题检测的不准确。针对此问题普遍的解决方法是通过主题差异性确定 K 取值。但微博平台用户量大，对每个用户都通过此办法测定 K 取值会带来算法复杂度大、兴趣挖掘效率低的问题。

b）由于微博文本具有长度较短、内容形式丰富多样、语言不规范、网络流行语多等特点，使得提取特征后用于兴趣建模的用户语料的词频信息及上下文信息严重缺乏，影响微博文本的主题建模效果。

c）没有考虑到用户的兴趣随时间变化的特点。

本文将在 LDA 模型的基础上提出 TCID-MUIM 挖掘方法，通过在 LDA 建模前后引入同义词合并策略、二次 Single-pass 不完全聚类算法、基于时间因子的主题矩阵压缩方法解决上述问题，用于挖掘用户兴趣主题。

2.1 同义词合并策略

对于微博平台的用户兴趣挖掘，首要步骤就是对微博文本进行去噪处理、分词处理、停用词处理、去除低频词及单字词等预处理操作，将每条微博文本表示为特征词的集合。然而微博文本较短小且长尾特征明显，去除低频词会使建模过程中词频信息不足的问题进一步加剧。因此，本文通过基于《同义词词林》的同义词合并策略，将用户微博文本中存在的低频词合并到高频词、单字词合并到多字词，对部分低频词及单字词进行合理利用。

《同义词词林》[11]是梅家驹先生等人按意义进行编排一部类义词典，包含一个词语的同义词以及其广义的相关词，其扩展版收录词语 7 万余条。词林以词义为主，兼顾词类，把词语划分为 5 级结构，不同等级通过不同的编码表示。具体标记如表 1 所示。

表 1 词语编码

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第1级	第2级	第3级		第4级	第5级		

表中的编码位是按照从左到右的顺序排列。在第五级（6、7 位）中，每个分类里词语数量已较少，难以进一步进行分割，称为原子词群或原子节点。而编码位的第 8 位仅是标记位，具有“=”、“#”、“@”三种标记，分别代表词行“同义”“同类”“独立”。通过词林中标记为“=”的同义词行可以实现对词汇的同义词替换。

基于《同义词词林》的同义词合并策略的具体步骤如算法 1 所示。

算法 1 同义词合并策略

输入：用户 u 的微博文本集 $D_u=\{d_{u1},d_{u2},\cdots,d_{uM}\}$。

输出：处理后的微博文本集 $W_u=\{w_{u1},w_{u2},\cdots,w_{uM}\}$。

步骤 1 对用户微博文本集 D_u 进行预处理，建立用户词库 V_u，并根据预处理后文本集 D_u 中词汇的出现次数统计用户词库中词汇的词频。

步骤 2 设定阈值 υ，将用户词库 V_u 中词汇词频大于等于 υ 的词放入高频词表，低于 υ 的词放入低频词表；并将 V_u 中出现的单字词放入低频词表。

步骤 3 根据《同义词词林》中的词语编码位，将第 8 位中符号位为“#”的同类词和符号位为“@”的独立词词行剔除，只保留符号位为“=”的同义词词行。

步骤 4 根据高频词表中词汇词频从高到低的顺序为词汇进行统一编号 $G=\{1,2,\cdots,i,\cdots,n\}$，从编号为 1 的高频词开始，将高频词与《同义词词林》进行匹配。若高频词 i 与词林匹配成功，则将匹配成功的词行作为高频词 i 的背景词行；否则，跳到下一个高频词，直到所有高频词与词林匹配恰好一次。

步骤 5 从编号为 1 的高频词开始，将低频词按词汇词频从高到低的编号次序与高频词背景词行中的词汇进行匹配，匹配成功，则赋予低频词与该高频词相同的词汇编号，并在中间词表中加入该词，从低频词表中删除该词，直到所有高频词的背景词行都与低频词表中的词汇匹配恰好一次。

步骤 6 将经过步骤 5 后低频词表中仍存在的词汇加入用户停用词表，使用用户停用词表对经过第一次预处理的用户微博文本集再去一次停用词。

步骤 7 利用高频词表及中间词表中词及编号的对应关系，将经过步骤 6 处理的用户微博文本集中的词汇变换为编号；再利用高频词表中词及编号的对应关系，通过反变换将编号转换为词汇。实现从低频词到高频词、单字词到多字词的同义词合并。

步骤 8 输出同义词合并后的微博文本集 W_u。

2.2 二次 Single-pass 不完全聚类算法

Single-pass 聚类算法又称单通道法或单遍法，是一种简单的增量聚类算法。通过数据对象的出现次序依次进行聚类处理，根据相似度值进行匹配。若相似度值大于事先设定的阈值，则将数据对象归入该类簇；否则将该数据对象作为一个新的聚类簇。

Single-pass 聚类可随着文本数量的增多而动态地变化，适用于用户微博列表不断增多的微博聚类。但具有明显的次序依赖的问题，可能导致较早完成遍历的微博因为与之前得到的话题的相似度略低于阈值而被重新创建了新的话题，影响聚类效果。因此，本文采用二次 Single-pass 算法对用户微博文本进行聚类。同时，考虑到用户兴趣在一定时间内具有一定的内聚性，在一定时间段内发布或转发的微博文本可能属于同一类别，兴趣方向变化不大，因此第一次 Single-pass 聚类时仅通过时间标记将后续输入的文本与前面时间标记最近的 m 个簇进行相似性比较。考虑到用户微博列表中的微博文本主题方向较杂乱，因此第二次 Single-pass 聚类时，不对第一次 Single-pass 聚类的

微博孤点进行处理，使用不完全聚类[12]的方法进行聚类。

由于微博文本特征词较少，所以选择 Jaccard 相似系数作为聚类算法中相似度的计算基础。对于给定的词集 A、B，Jaccard 相似系数的计算公式为

$$sim_J(A,B)=\frac{A\cap B}{A\cup B} \tag{1}$$

若簇的微博集合为 $c_K=\{w_1,w_2,\cdots,w_k\}$，其对应词集为 $A_{c_K}=\{A_{w_1},A_{w_2},\cdots,A_{w_k}\}$，$A_{w_i}$ 为 w_i 对应的词集，其中 $i=1,2,\cdots,k$。目标微博 w_m 对应的词集为 B_{w_m}，则第一次 Single-pass 聚类时，微博与簇间的相似度计算公式可定义如下：

$$sim\left(A_{c_K},B_{w_m}\right)=\frac{1}{k}\times\sum_{1\le i\le k}\left\{sim_J\left(A_{w_i},B_{w_m}\right)\right\} \tag{2}$$

在式（2）中，通过计算 B_{w_m} 与微博簇 A_{c_K} 的平均相似度值来衡量目标微博 w_m 与簇 c_K 间的相似度。

相同地，簇 $c_K=\{w_1,w_2,\cdots,w_k\}$ 与簇 $c_L=\{w_1,w_2,\cdots,w_l\}$ 间的相似度计算公式定义如下：

$$sim\left(A_{c_K},A_{c_L}\right)=\frac{1}{k\times l}\times\sum_{\substack{1\le i\le k\\1\le j\le l}}\left\{sim_J\left(A_{w_i},A_{w_j}\right)\right\} \tag{3}$$

二次 Single-pass 不完全聚类算法的具体步骤如算法 2 所示。

算法 2　二次 Single-pass 不完全聚类算法

输入：用户 u 的微博文本集 $W_u=\left\{w_{u1},w_{u2},\ldots,w_{uM}\right\}$，阈值 τ，ξ，比较次数 m。

输出：簇文本及微博孤点的集合 $W_u^*=\left\{w_{u1}^*,w_{u2}^*,\ldots,w_{uQ}^*\right\}$ 和 K 值。

步骤 1　将微博文本 w_{ui} 的发布时间设置为其时间标记 t_{ui}，其中 $i=1,2,\cdots,M$。按时间标记 $t_{u1},t_{u2},\ldots,t_{uM}$ 的顺序输入微博文本 $w_{u1},w_{u2},\ldots,w_{uM}$，其中 t_{uM} 为最新一条微博的发布时间。针对 W_u 进行处理得到对应的词集 $A_{W_u}=\{A_{w_{u1}},A_{w_{u2}},\ldots,A_{w_{uM}}\}$。

步骤 2　当 $i=1$，簇数 $J=1$ 时，将时间标记为 t_{u1} 的微博文本 w_{u1} 作为第一个聚类簇 c_1，即 $c_1=\{w_{u1}\}$，执行 $i=i+1$，转到步骤 3。

步骤 3　当 $i\le M$ 时，若 $J>m$，则根据式（2）计算 $a_j=\max\limits_{J-m\le j\le J}\left\{sim\left(A_{c_j},A_{w_{ui}}\right)\right\}$；否则根据式（2）计算 $a_j=\max\limits_{1\le j\le J}\left\{sim\left(A_{c_j},A_{w_{ui}}\right)\right\}$。若 $a_j>\tau$，则转到步骤 3.1；否则转到步骤 3.2。

步骤 3.1　更新 $c_j=c_j\cup w_{ui}$，用微博 w_{ui} 的时间 t_{ui} 标记更新簇 c_j 的时间标记，执行 $i=i+1$，转到步骤 3。

步骤 3.2　执行 $J=J+1$，建立新簇 $c_J=\{w_{ui}\}$，将时间标记为 t_{ui} 的微博文本 w_{ui} 作为新簇 c_J 的时间标记，执行 $i=i+1$，转到步骤 3。

步骤 4　执行 $C=\{c_j:|c_j|\ge 2,\ 1\le j\le J\}$ 以及 $C_1=\{c_j:|c_j|=1,\ 1\le j\le J\}$，转到步骤 5。

步骤 5　将簇族 C 重新标记为 $C^*=\{c_s^*\}_{s=1}^{|C|}$，按时间标记 $t_1,t_2,\ldots,t_{|C|}$ 的顺序输入微博簇族 C^*，其中 t_1 为离当前时间最近的时间标记。将 C^* 对应的词集记为 $A_{C^*}=\{A_{c_s^*}\}_{s=1}^{|C|}$，这里的 $A_{c_s^*}$ 为 c_s^* 对应的词集。

步骤 6　当 $s=1$，簇数 $I=1$ 时，将时间标记为 t_1 的簇 c_1^* 作为第一个聚类簇 c_1^*，执行 $s=s+1$，转到步骤 7。

步骤 7　当 $s\le|C^*|$ 时，根据式（3）计算 $b_r=\max\limits_{1\le r\le I}\left\{sim\left(A_{c_r^*},A_{c_s^*}\right)\right\}$，若 $b_r>\tau$，则转到步骤 7.1；否则，转到步骤 7.2。否则，转到步骤 8。

步骤 7.1　更新 $c_r^*=c_r^*\cup c_s^*$，执行 $s=s+1$，转到步骤 7。

步骤 7.2　执行 $I=I+1$，更新簇 $c_I^*=c_s^*$，执行 $s=s+1$，转到步骤 7。

步骤 8　将步骤 7 的结果按照簇内微博数的数目从大到小排序，并将排序结果标记为 $C^{**}=\{c_1^{**},c_2^{**},\cdots,c_I^{**}\}$，将 C^{**} 中的 I 个微博簇分别合并为 I 个簇文本 $w_1^*,w_2^*,\ldots,w_I^*$，簇的时间标记作为该文本的时间标记。

步骤 9　初始时 K=1，计算 $\varepsilon_1=|c_1^{**}|/M$，若 $\varepsilon_1\ge\xi$，则输出 K；否则执行 $K=K+1$，转到步骤 10。

步骤 10　当 $K\le I$ 时，计算 $\varepsilon_K=\sum_{1\le k\le K}|c_k^{**}|/M$，若 $\varepsilon_K<\xi$，则执行 $K=K+1$；否则，输出 K。

步骤 11　执行 $W_u^*=C_1\cup\{w_1^*,w_2^*,\ldots,w_I^*\}$，将 W_u^* 的结果重新标记为 $W_u^*=\left\{w_{u1}^*,w_{u2}^*,\ldots,w_{uQ}^*\right\}$，此时 $Q=|C_1|+I$。

注：在步骤 1 中，A_{W_u} 中的 $A_{w_{ui}}$ 为 w_{ui} 对应的词集，其中 $i=1,2,\cdots,M$。步骤 2～3.2 完成第一次 Single-Pass 聚类。在步骤 3 中，a_j 的下标 j 表示 $\max\limits_{1\le j\le J}\left\{sim\left(A_{c_j},A_{w_{ui}}\right)\right\}$ 或 $\max\limits_{J-m\le j\le J}\left\{sim\left(A_{c_j},A_{w_{ui}}\right)\right\}$ 取得最大值的下标。在步骤 4 中，C 表示微博数大于等于 2 的微博簇族，C_1 表示簇内微博数等于 1 的孤点簇族。步骤 6～7.2 完成第二次 Single-Pass 聚类。在步骤 7 中，b_r 中的下标 r 表示 $\max\limits_{1\le r\le I}\left\{sim\left(A_{c_s^*},A_{c_r^*}\right)\right\}$ 取得最大值的下标。$|C|$，$|C^*|$、$|C_1|$ 分别表示集合 C、集合 C^* 和 C_1 中的元素个数。在步骤 9～10 中，按照文献[13]描述的“微博短文本聚类后其聚类结果具有长尾分布的特征”的思想，根据长尾分布的一般取值原则可将阈值 ξ 设置为 0.20；通过 ξ 计算得到的 K 值即为 W_u^* 中的大簇数目。

2.3　基于时间因子的主题矩阵压缩方法

用户的兴趣可以分为长期兴趣与短期兴趣。长期兴趣是指用户长时间保持的兴趣偏好，它不会随时间的流逝而造成大的变化；而短期兴趣则是指因特定原因导致的用户兴趣短期偏移，其特点是用户在一定时期内会大量关注与该兴趣相关的微博信息，但在一段时间以后，用户对该兴趣相关微博信息的关注会迅速衰减，甚至不再关注。

用户兴趣变化体现在用户发布的历史微博主题随时间的变化中，变化过程与 Ebbinghaus 遗忘曲线[13]遵循同样的规律，在已有的兴趣范围内，新的兴趣不断诞生的同时，旧的兴趣也在不断地衰减，甚至遗忘。根据这一特点，本文提出了基于时间

因子的主题矩阵压缩方法，根据记忆函数给用户近期微博的主题分布赋予较高的记忆值（memory value），根据记忆值压缩微博—主题矩阵，获取用户—主题矩阵，得到用户的兴趣主题分布。使得用户微博列表中近期发布的微博对用户兴趣的影响越大，尽可能减少短期兴趣对用户兴趣的影响，使获取的用户兴趣主题分布更贴近用户当前的兴趣偏好，并将微博—主题矩阵 Θ 压缩为易于理解的用户—主题矩阵 θ_u。

设用户 u 最近一条微博的发布时间为 t_{ur}，若 $W_u^*=\{w_{u1}^*,w_{u2}^*,\cdots,w_{uQ}^*\}$ 中第 m 个文本 w_{um}^* 的时间标记为 t_{um}，则兴趣建模时其记忆值 $mv(w_{um}^*,t_{ur})$ 定义如下：

$$mv(w_{um}^*,t_{ur})=2^{-\lambda(t_{ur}-t_{um})} \tag{4}$$

其中：$\lambda>0$，λ 越大，历史数据的重要性随时间降低的越快。

根据式（4），可以利用 W_u^* 中每个文本对应的时间标记求取微博—主题矩阵 $\Theta_u=\{\theta_m\}_{m=1}^{Q}$ 每一行（微博的主题分布）对应的记忆值。

$$\begin{array}{ccc} \text{memory value} & & \begin{array}{cccc} T_{u1} & T_{u2} & \cdots & T_{uK} \end{array} \\ mv\left(w_{u1}^*,t_{ur}\right) & w_{u1}^* & \\ mv\left(w_{u2}^*,t_{ur}\right) & w_{u2}^* & \\ \cdots & \cdots & \\ mv\left(w_{uQ}^*,t_{ur}\right) & w_{uQ}^* & \end{array} \left[\begin{array}{cccc} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ p_{Q1} & p_{Q3} & \cdots & p_{QK} \end{array}\right]$$

根据记忆值以及主题 T_{uk} 对应的每一条微博维上的概率分布值，可以求得主题 T_{uk} 在用户—主题分布中的概率值 $P(T_{uk})$。计算公式如下：

$$P(T_{uk})=\frac{\sum_{m=1}^{Q}\left[mv\left(w_{um}^*,t_{ur}\right)\times p_{mk}\right]}{\sum_{k=1}^{K}\sum_{m=1}^{Q}\left[mv\left(w_{um}^*,t_{ur}\right)\times p_{mk}\right]},\quad k=1,2,\cdots,K \tag{5}$$

根据式(5)可以得到用户最终的兴趣主题分布 $\theta_u=(P(T_{u1}),P(T_{u2}),\cdots,P(T_{uK}))$。

2.4　TCID-MUIM 挖掘方法

在中文微博平台，用户历史微博列表主要包括转发微博与原创微博两种类型的微博。转发微博由转发部分和原创部分组成。原创部分是自己对转发内容的附加内容。同时，微博中有一类特殊的功能符号——话题标签，它是微博同一类话题的标志，同一用户历史微博中具有相同话题标签的微博通常话题相关性加强。针对这一特点，结合 LDA 模型及上述同义词合并策略、二次 Single-pass 不完全聚类算法、基于时间因子的主题矩阵压缩方法，将 TCID-MUIM 挖掘方法的具体步骤如算法 3 所示。

算法 3　TCID-MUIM 挖掘算法

输入：用户 u 的微博文本集 $D_u=\{d_{u1},d_{u2},\cdots,d_{uM}\}$。

输出：用户—主题分布 θ_u、主题-词汇矩阵 Φ_u。

步骤 1　对微博进行第一次初始聚类：将转发微博与转发时发布的原创微博合并；根据提取的话题标签，将含有同一标签的微博文本合并。

步骤 2　根据算法 1 同义词合并策略，对步骤 1 处理后的微博文本进行同义词合并处理。

步骤 3　根据算法 2 二次 Single-pass 不完全聚类算法，对步骤 2 处理后的微博文本进行聚类，获取簇文本及微博孤点的集合 W_u^*、K 值。

步骤 4　将 W_u^* 作为 LDA 建模的语料，K 值设为 LDA 建模的主题数。通过 LDA 模型获取用户 u 的微博-主题矩阵 Θ_u、主题-词汇矩阵 Φ_u。

步骤 5　根据 3.3 节中描述的基于时间因子的主题矩阵压缩方法，利用微博文本的时间标记计算记忆值，将微博—主题矩阵 Θ_u 压缩为用户—主题分布 θ_u。

步骤 6　输出用户—主题分布 θ_u，主题—词汇矩阵 Φ_u。

3　实验

3.1　实验设置

目前，在国内中文微博平台的相关研究领域，还没有用于评测的标准数据集。本文通过 firefox 浏览器及爬虫插件 datascraper、metastudio 获取实验数据，采集了新浪微博平台 1 356 位用户，共计 91 万余条微博数据。根据实验需要，过滤掉其中历史微博数小于 150 的用户，将保留的 623 位用户的微博数据按发布时间分为两部分，其中发布时间较近的前 50 条微博作为测试集，其余部分作为兴趣模型训练集。提取博文内容中的话题标签属性后，对数据进行去噪处理、分词处理（ICTCLAS 分词系统[14]）、去停用词处理，将处理后的数据用于实验。

LDA 建模参数设置为 alpha=0.1，beta=0.1，iter_times=100；TCID-MUIM 方法中涉及的参数，分别将阈值参数设置为 $\upsilon=3$，$\tau=0.25$，$\xi=0.20$，比较次数设置为 $m=100$。

3.2　主题有效性

微博主题挖掘的目标是从海量信息中挖掘出能代表用户兴趣的兴趣主题，并匹配相关性高的词汇描述主题，词汇与主题的匹配程度越高，则认为主题有效性越高。为了验证 TCID-MUIM 挖掘方法的有效性，实验另设置了以下两种方法进行对比：

a）文本合并（text merge，TM）。将用户的所有历史微博合并为一个文本，使用 LDA 模型对合并后的文本建模。

b）传统方法（conventional method，CM）。不对文档进行同义词合并、聚类处理，直接使用 LDA 模型对用户微博文本集建模。

三种方法的建模主题数 K 设定为 TCID-MUIM 挖掘方法中确定的大簇数目。表 2 显示了三种方法下用户兴趣挖掘的实验结果。限于篇幅，表中只列出 clueid 为 121365465 的用户根据三种方法所获取的 3 个兴趣主题，每个主题由概率最大的前 10 个单词表示。通过观察主题所属词汇并比对数据可以看出，这 3 个主题分别描述的是与医患关系、法治、教育相关的主题；虽然三种方法所挖掘的主题都能在一定程度上表达用户兴趣，但 TCID-MUIM 挖掘方法挖掘到的主题与其对应的关键词汇匹配准确率较高，主题集中性更强。例如在 Topic 1st 中，TCID-MUIM 方法下的关键词都与医患关系有很强的相关性，而 TM 与 CM 方法下的关键词中却存在法治、政府等相关性较低的词汇。

表 2　主题有效性对比

TOPIC 1st						TOPIC 2nd						TOPIC 3rd					
TCID-MUIM		TM		CM		TCID-MUIM		TM		CM		TCID-MUIM		TM		CM	
医院	0.0174	治疗	0.0058	医院	0.0199	律师	0.0288	违法	0.0048	律师	0.0216	学生	0.0273	教师	0.0211	学生	0.0234
患者	0.0059	法治	0.0048	北京	0.0107	取保候审	0.0087	反抗	0.0043	法院	0.0108	学校	0.0098	教授	0.0195	学校	0.0113
医生	0.0059	医院	0.0043	患者	0.0080	法院	0.0077	犯罪	0.0043	国家	0.0089	研究	0.0091	大学	0.0120	大学	0.0106
百度	0.0059	死亡	0.0039	政府	0.0067	组织	0.0053	律师	0.0037	公开	0.0083	教育	0.0082	学院	0.0102	校方	0.0082
医疗	0.0056	魏则西	0.0037	百度	0.0062	人权	0.0048	行政	0.0037	国际	0.0073	大学	0.0078	学生	0.0085	校长	0.0067
魏则西	0.0051	北京	0.0037	莆田	0.0053	涉嫌	0.0048	法官	0.0037	判决	0.0070	同学	0.0069	教育	0.0082	政府	0.0063
莆田	0.0051	身体	0.0032	魏则西	0.0053	罪名	0.0044	人权	0.0037	政府	0.0051	小学	0.0062	校长	0.0079	老师	0.0059
良心	0.0043	良心	0.0032	医疗	0.0049	羁押	0.0044	披露	0.0037	庭审	0.0051	老师	0.0056	政府	0.0069	教育	0.0059
家属	0.0039	记者	0.0032	医生	0.0049	庭审	0.0039	酒店	0.0037	要求	0.0048	文化	0.0052	北京	0.0067	研究	0.0051
北京	0.0034	医疗	0.0032	记者	0.0045	判决	0.0039	仲裁	0.0032	案件	0.0048	实验	0.0049	文化	0.0061	文化	0.0043

3.3　主题差异性

主题差异性是指主题模型生成的主题分布间的差异程度。主题模型的思想是，建模后主题间的差异度越大、相似度越小，则认为主题越有代表性，模型的建模效果越好。基于此，本文通过距离测度的方法度量主题模型建模后获取的主题分布间的差异性，以此衡量 TCID-MUIM 建模方法的性能。具体方法是：首先，通过度量两个概率分布间差异程度的 jensen-shannon(JS) 距离计算模型生成的各个主题间的差异度；然后，根据上一步骤的计算结果计算主题平均差异度，并将平均差异度作为衡量模型主题差异性的指标。

JS 距离计算公式如下：

$$D_{JS}(\varphi_i,\varphi_j)=\frac{1}{2}\left[D_{KL}(\varphi_i,\frac{\varphi_i+\varphi_j}{2})+D_{KL}(\varphi_j,\frac{\varphi_i+\varphi_j}{2})\right] \quad (6)$$

$$D_{KL}(\varphi_i,\frac{\varphi_i+\varphi_j}{2})=\sum_{n=1}^{n} p_{in}\ln\frac{2p_{in}}{p_{in}+p_{jn}} \quad (7)$$

其中：$\varphi_i=(p_{i1},p_{i2},\cdots,p_{in})$ 与 $\varphi_j=(p_{j1},p_{j2},\cdots,p_{jn})$ 分别是两个主题 T_i 与 T_j 的词概率分布。

本实验通过上述方法度量 TCID-MUIM、TM 和 CM 三种建模方法的主题差异性。三种建模方法中两两主题之间的 JS 平均距离比较如表 3 所示。可以看出，三种建模方法中 TM 方法的主题差异性（JS 值）相较于其他两种建模方法较低，这可能是由于强行抹去文本边界的原因；而 TCID-MUIM 方法在三种建模方法中的 JS 值相对较大，具有更好的主题差异性，说明了对微博进行同义词合并及不完全聚类能够帮助主题模型发现更具代表性的主题。

表 3　三种建模方法 JS 距离比较

建模方法	JS 距离
TCID	0.368191
TM	0.297962
CM	0.353599

3.4　兴趣挖掘效果对比

为了判断 TCID-MUIM 挖掘用户兴趣的准确度，本文将用户最新发布的 50 条微博作为测量数据，其余部分的微博用作用户兴趣主题建模数据，对比 TCID-MUIM、CM、TM 三种挖掘方法的主题建模效果。评价指标选择预测准确率、漏检率以及本文自定义的概率准确率。具体的计算公式如下：

$$P=\frac{N_C}{N_T} \quad (8)$$

其中：P 表示预测准确率；N_C 表示测量数据中属于用户兴趣主题的微博数；N_T 表示测量数据中的微博总数。

$$P_{MISS}=\frac{N_M}{N_T} \quad (9)$$

其中：P_{MISS} 表示漏检率；N_M 表示测量数据中与用户兴趣主题不相关的微博数；N_T 表示测量数据中的微博总数。

$$P_{AC}=\sum_{i=1}^{k} c_i\times p_{u,i} \quad (10)$$

其中：P_{AC} 表示概率准确率；$p_{u,i}$ 是用户兴趣主题分布 $\theta_u=(p_{u,1},p_{u,2},\cdots,p_{u,k})$ 中第 i 个兴趣主题的分布概率；c_i 是测量数据中属于第 i 个兴趣主题的微博数目。为了在图表中更直观的展示三种方法下的 P_{AC}，选择将三种方法下的 P_{AC} 值归一化处理，并放大一倍。

图 4 显示了在采用 LDA 模型进行用户兴趣建模的条件下，TCID-MUIM、CM、TM 三种不同用户兴趣挖掘方法进行兴趣主题挖掘的实验对比效果。观察可发现，CM 与 TM 方法相比，TM 方法概率准确率高于 CM，但预测准确率低于 CM，这可能是由于 TM 方法将所有微博文本合为一个文档，导致主题集中性强，单一主题概率高，但主题区分度较低，不能更大范围地涵盖用户兴趣偏好所致；而 TCID-MUIM 方法因将文本语料进行了不完全聚类处理，保证了主题的集中性，同时考虑了兴趣衰减问题，保证了主题与用户实时兴趣的贴合性，所以 TCID-MUIM 与 CM、TM 方法相比在三种评价指标下都具有明显优势。

4　结束语

本文针对直接使用 LDA 模型挖掘用户兴趣时存在的微博文本长度较短、语义信息缺乏影响主题建模效果以及没有考虑用户兴趣随时间变化的问题，提出了基于文本聚类与兴趣衰减的微博用户兴趣挖掘算法 TCID-MUIM，通过同义词合并策略、二次 Single-pass 不完全聚类算法、LDA 模型建模方法以及基于

时间因子的主题矩阵压缩方法挖掘用户兴趣主题。在真实微博数据集上对 TCID-MUIM 相关实验验证进行实验，实验结果表明通过 TCID-MUIM 方法挖掘的用户兴趣主题与 TM、CM 方法相比具有更好的主题区分度，且更贴合用户的真实兴趣偏好。下一步工作需要解决的问题是用户兴趣挖掘中的冷启动问题，考虑是否可以利用用户其他网站上的历史数据信息构建用户兴趣模型。

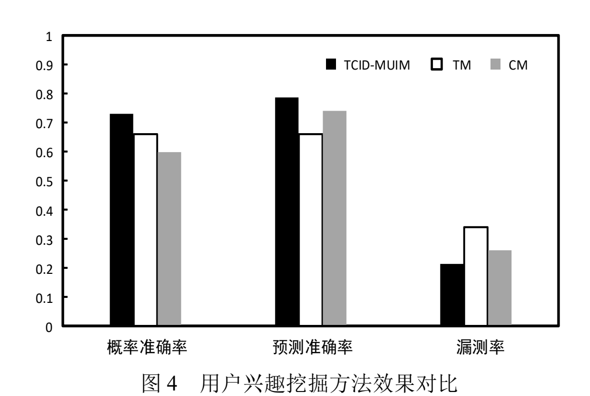


图 4　用户兴趣挖掘方法效果对比

参考文献：

[1] China Internet Network Information Center. The 39th statistical report on Internet development in China [EB/OL]. (2017-01-22) . http://www. cnnic. net. cn/hlw fzyj/hlwxzbg/hlwtjbg/201701/t20170122_66437. htm.

[2] 企鹅智酷. 2016 微博用户研究报告 [EB/OL]. (2016-09-09) . http://tech. qq. com/original/archives/a124. html.

[3] Chen J, Nairn R, Nelson L, *et al.* Short and tweet: experiments on recommending content from information streams [C]// Proc of International Conference on Human Factors in Computing Systems. 2010: 1185-1194.

[4] Abel F, Gao Q, Houben G J, *et al.* Twitter-based user modeling for news recommendations [C]// Proc of the 23rd International Joint Conference on Artificial Intelligence. [S. l.] : AAAI Press, 2013: 2962-2966.

[5] Hong L, Davison B D. Empirical study of topic modeling in twitter [C]// Proc of the 1st Workshop on Social Media Analytics. [S. l.] : ACM Press, 2010: 80-88.

[6] 刘红兵，李文坤，张仰森. 基于 LDA 模型和多层聚类的微博话题检测 [J]. 计算机技术与发展, 2016, 26 (6): 25-30.

[7] 唐晓波，房小可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究 [J]. 情报理论与实践, 2013, 36 (8): 85-90.

[8] 张培晶，宋蕾. 基于 LDA 的微博文本主题建模方法研究述评 [J]. 图书情报工作, 2012, 56 (24): 120-126.

[9] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[10] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 (1): 5228.

[11] 梅家驹. 同义词词林 [M]. 上海：上海辞书出版社, 1983.

[12] 彭泽映，俞晓明，许洪波，等. 大规模短文本的不完全聚类 [J]. 中文信息学报, 2011, 25 (1): 54-59.

[13] 于洪，李转运. 基于遗忘曲线的协同过滤推荐算法 [J]. 南京大学学报：自然科学版, 2010, 46 (5): 520-527.

[14] Zhang H P, Yu H K, Xiong D Y, *et al.* HHMM-based Chinese lexical analyzer ICTCLAS [C]// Proc of the 2nd SIGHAN Workshop on Chinese Language Processing. 2003: 184-187.